DRAGEN v4.1.23 Software Release Notes

Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform v4.1.23.

Changes are relative to DRAGEN™ v4.1.7. If you are upgrading from a major version prior to DRAGEN™ v4.1, please review the release notes for a list of features and bug fixes introduced in subsequent versions.

DRAGEN™ Installers, User Guide and Release Notes are available here:
https://support.illumina.com/sequencing/sequencing_software/dragen-bio-it-platform.html

The software package includes downloadable installers for Phase 3 and Phase 4 on-site servers:
- DRAGEN™ SW for x86 Centos 7 - dragen-4.1.23-9.el7.x86_64.run
- DRAGEN™ SW for x86 Oracle 8 - dragen-4.1.23-9.el8.x86_64.run

The following configurations containing DRAGEN™ 4.1.23 are also available on request:
- Centos 7 and Oracle 8 Amazon Machine Images (AMI) for f1 instances, available in 12 regions
- Centos 7 Microsoft Azure Image (VM) available in West US 2
- Centos 7 and Oracle 8 RPM packages for use with Amazon Web Services (AWS) f1 instances, for customer generated AMIs or customer generated docker images
- DRAGEN™ Kernel drivers for el7 and el8, for use with customer generated AMIs or QuickStart
- Pre-built docker images with Centos 7 and Oracle 8 for on-site, AWS usage
- Pre-built docker image with Centos 7 for Microsoft Azure cloud usage

Deprecated platforms:
- Support for DRAGEN™ Server v1 FPGA cards have been deprecated since DRAGEN™ v3.10
- Support for Ubuntu has been deprecated since DRAGEN™ v3.9
- Support for x86 CentOS 6 has been deprecated since DRAGEN™ v3.8

Contents

Overview

Below is a summary of the changes included in v4.1.23. This is a minor update to DRAGEN™ v4.1:

- Various bug fixes.
- Robustness and runtime improvements for NovaSeq-X on-instrument analysis.
- Added sample sheet options for QC metrics and BAM tags for NovaSeq-X on-instrument analysis.
- New workflows for NovaSeq-X analysis: Somatic WGS T-Only and DNA Methylation.
- Added CYP21A2 and GBA callers to the Germline WGS workflow.
- The changes and/or fixes listed in the sections below apply to server, cloud, and on-instrument workflows.

Updates and Fixes

NovaSeq-X on-instrument

- Various improvements have been made to address on-instrument system robustness, memory usage and runtime.
- Enable Star Allele, CYP21A2, GBA callers in the WGS Germline workflow when all callers are enabled.
- New workflows for Somatic WGS T-Only, and DNA Methylation analysis.
- Support enabling of multiple QC coverage metrics from Sample Sheet.
- Support specification of BAM tags in the Sample Sheet Data section.
- Added a disk space watchdog to abort ongoing analysis runs and clean up files when the system detects a full disk.

BCL Conversion

- Combined index collision checking remains enabled by default for all lanes.

DRAGEN™ version	Index collision check behavior
3.9.x	Relaxed by default. No option to change. Matches bcl2fastq2
3.10.x and 4.0.x	Strict by default. No option to change.
4.1.5	Strict by default. New option `CombinedIndexCollisionCheck` introduced to optionally relax the strictness
4.1.7, 4.1.23 and 4.2.x	Relaxed by default. Remove `CombinedIndexCollisionCheck` option, add new `IndependentIndexCollisionCheck` option to allow optional strict checking. Default matches bcl2fastq2

- Fixed behavior so that BCL does not output Undetermined FASTQs when `--bcl-only-matched-reads` is enabled.
- Fixed issue where the ORA outputs were generated in the root output directory, instead of inside the Sample_Project and Sample_Name directories like FASTQ.gz files.
- Prevent using `--no-sample-sheet` option and the `--fastq-compression-format=dragen` (or dragen-interleaved) option together.
- Fixed an ORA error when FASTQ.gz files for a sample-read were not generated if the whole read was masked out as N or U.
- Fixed a performance issue with very high sample counts, such as 150K samples.

- Fixed an issue where customers with high CPU core count systems have reduced BCL performance due to a thread limit, since v3.10.
- Fix for BCL crash when `--no-sample-sheet true` and 0 indexes supplied.
- Fixed an issue where a large number of demultiplex cycles (~26) caused a hang and crash
- Fixed an issue where specifying an invalid lane in 'bcl-only-lane' parameter errored out without providing an error message.
- Support for values 'true' and 'false' for sample sheet settings 'TrimUMI' and 'CreateFastqForIndexReads', in addition to '1' and '0'
- Fixed an issue that failed to catch an invalid integer parameter for TrimUMI (must be 0 or 1 if an integer)
- Fixed a crash issue with i5 index lengths greater than 18

Other Bug Fixes

- Fixed an issue where scATAC produces empty outputs (barcode list, matrix) when using combinatorial barcodes.
- Fix for incorrect HLA output when running back-to-back. Runs with HLA enabled could end up with a partially incorrect reference loaded for mapping after the run is finished. With back-to-back runs, and without reference re-loading, the first mapping stage of subsequent runs fail to re-load the reference and lead to potentially incorrect mapper outputs. This fix results in a minor accuracy regression
- Map/Align:
    - Fixed a mapper issue where discordant or half-mapped pairs' primary alignments are linking to each other's PNEXT, causing POS/PNEXT mismatches which fails Picard ValidateSamFile QC check. The issue is only present when specific non-default PE overhang trimming mapper options are used.
    - Improve the insert length (mean, median, standard deviation) reported in the MAPPING/ALIGNING SUMMRY metrics, to contain the full sample statistics, instead of statistics over a small initial interval. This update makes the metric accurate for the whole sample.
    - Fixed an issue where Somatic T/N end-to-end runtime doubles when map/align output (BAM/CRAM) is enabled, when starting from FASTQ/ORA and enable map/align + VC.
- ORA compression:
    - Restore exact filenames for decompressed interleaved FASTQs that are ORA compressed. Previously, they were not the same as the original file names.
    - Option `--ora-original-name` to restore original filename if present when decompressing (default false)
    - Generate a fastq_list.csv for BCL to ORA
    - Allow `--force` option to overwrite ORA files when compressing.
    - Improve robustness when using the settings `--ora-threads-per-file` and `--ora-parallel-files`.

New Targeted Caller

- **CYP21A2 variant detection from WGS data**
    - Pathogenic variants in the CYP21A2 gene can cause 21-Hydroxylase-Deficient Congenital Adrenal Hyperplasia, an autosomal recessive disease.
    - Reads mapping to either CYP21A2 or CYP21A1P are used to detect variants in homology regions and read phasing is used to detect any recombination events between the two genes.

- Usage:
  - To enable CYP21A2 caller: `--enable-cyp21a2=true`

    Example fields from `*targeted.json` output file:

```
{
    "sample": "HG01801",
    "cyp21a2": {
            "totalCopyNumber": 3,
            "recombinantHaplotypes": [
                "NM_000500.9:c.518T>A",
                ""
     ...
}
```

    * *Supported references: hs37d5, hg19, hg38*
- Highly concordant with orthogonal methods
  - Validation samples used:
    - 14 affected samples from Radboudumc (Long-range PCR, MLPA)
    - 66 unaffected 1KGP samples (some carriers)
    - 4 Coriell samples (2 affected and 2 carriers)

Benchmark set	Total samples	Concordant
Radboudumc	14	13 (**92.86%**)
1KGP	66	65 (**98.48%**)
Coriell	4	4 (**100%**)
Total	84	82 (**97.62%**)

Known Issues

Known issues of the DRAGEN™ v4.1.23 release

Component	Issue ID	Summary	Remedy/Workaround
Alignment	DRAGEN-23757	Initial insert size estimates can be significantly inaccurate for a sample when there are sequencing dropouts (no reads or coverage) over the first tiles of a flow cell, potentially impacting mapping accuracy.	The accurate full sample insert lengths are now calculated in the metrics. Larger insert sample size can be set via command line.
BCL	DRAGEN-18920	bcl-convert outputs different PF cluster YieldQ30 and QualityScoreSum metrics in the legacy stats file ConversionStats.xml as compared to bcl2fastq2.	No workaround. Fix planned for future version
BCL	DRAGEN-19103	BCL crashes in Robust mode when converting a single lane of an aggregated (bgzf) format, and the lane's filter file is missing or corrupted.	No conversion can be done without some P/F data from the filter file. Because the filter file is for the entire

			lane, the lane cannot be converted.
BCL	DRAGEN-26220	When using mixed indexing strategies, the index hopping counts metrics for Undetermined reads may differ slightly between bcl-convert and NovaSeq-X on-instrument	No workaround. Fix planned for future version
BCL	DRAGEN-26294	Bcl-convert fails to catch the error case where per-sample-settings is used and the same Sample_ID has multiple entries with a different set of expected output files for each entry due to fully masked reads. In this case, the set of files will be inconsistent and incomplete.	Rare corner case where the same Sample_ID may be used with, e.g. R1 masked out in one entry but not in another, in the same lane. Error will be caught in future versions.
CNV caller	DRAGEN-25042	Incorrect ploidy estimation on sample with large deletion, does not call the deletion	No workaround. Fix planned for future version
DNA Alignment	DRAGEN-22182	DRAGEN @RG header SAM/BAM/CRAM output not specifications compliant	For some input values of RGPL, DRAGEN generates RG line that is not SAM compliant and produce Bam that has compatibility issues with some 3rd party tools. The SAM header @RG sub field "PL" is listed as PL0
DRAGEN-MP	DRAGEN-28008	File IO error leading to segfault on NovaSeq-X	Rare issue seen on some systems due to bad I/O read from disk. No workaround.
Duplicate Marking	DRAGEN-23711	Very large samples can fail with the default dupmark-version=hash due to a system limitation. The system crash with "Assertion `pos < m_num_bits' failed.	Run with "dupmark-version=sort"
Gvcf Genotyper	DRAGEN-26325	Gvcf Genotyper truncates the names of contigs to the first colon. This leads to incorrect outputs for those contigs. Some references contain such HLA* contigs.	No workaround. Fix planned for future version
Gvcf Genotyper	DRAGEN-21091	When a site is missing in the input gVCF file for a sample and the site is output to the msVCF file, the genotype is coded as missing using '.' haploid. However, according to the VCF 4.2 specification missing genotype should be coded with '.' for each missing allele i.e './.' for a missing diploid genotype.	No workaround. Very rare occurrence and low impact. Fix planned for future version
Gvcf Genotyper	DRAGEN-21922	Some incorrect LPL and LAA values in msVCF	No workaround. Fix planned for future version
GVCF Genotyper	DRAGEN-26325	Contig name truncation on HLA* alt contigs	GG is truncating the names of contigs to the first colon. This

			leads to incorrect outputs for those contigs.
Hash Table	DRAGEN-26399	Hash table decompression error on some fasta files	Write the hash table uncompressed
HLA	DRAGEN-27177	HLA accuracy regression	Accuracy regression of up to 3% on some samples, compared to DRAGEN v4.0. Fix planned for future version
Imputation	DRAGEN-22549	Imputation end to end pipeline adds only the first chromosome name to VCF the header, leading to problems with downstream tools	Re-header the VCF using bcftools
Infra	DRAGEN-19988	A crash on Microsoft Azure cloud can leave the system in a bad state that requires intervention and prevents subsequent jobs form succeeding. "ERROR: xclRegRW: can't map CU: 0"	Known issue for which a solution is not available
Inputs	DRAGEN-26218	DRAGEN MA hangs due to extra reads in one end	If fastq input r2 has more reads than fastq input r1, then the systems hang.
Joint Genotyping	DRAGEN-19844	Pedigree Joint genotyping is up to 30% slower compared to v4.0	No workaround. Fix planned for future version
Joint Genotyping	DRAGEN-21909	Accuracy on denovo WGS joint genotyping changed, due to an ML qual adjustment made to improve NovaSeq-X indel performance	Planned FP/FN accuracy tradeoff for improved performance on NovaSeq-X data
Joint Genotyping	DRAGEN-25355	WES and WGS NNN generated more Fp in 4.0/4.1	Accuracy on denovo WGS joint genotyping changed, due to an ML qual adjustment made to improve NovaSeq-X indel performance
On-instrument	DRAGEN-25465	On-instrument NovaSeq-X runs can exceed a memory budget for BCL conversion and fail, when processing long reads such as 2x300	No workaround. Fix planned for future version
Paralog Caller	DRAGEN-25354	GBA variant NM_000157.4:c.84dup reported as NM_001005741.2:c.84dupG	GBA variant identifier is not the HGVS mane select like the other variant identifiers. Can still find the correct variant when searching with the provided identifier in clinvar.
QC Metrics	DRAGEN-24949	Inconsistent fastqc metrics between Local Server and Azure	FASTQC results differ on Azure vs Local run.

RNA Quant	DRAGEN-24824	RNA quant - SJ.saturation.txt differs with num-threads value	Minor difference when using different number of threads. Issue does not impact main quant output files and it is not reproducible when using a consistent number of threads across runs.
SNV VC	DRAGEN-22841	In rare cases, MNVs are wrong when the merging distance is greater than graph TLEN	No workaround. Fix planned for future version
SNV VC	DRAGEN-23630	An invalid alignment used to build the graph genome, leads to an incorrect allele frequency. Only one such instance has been found.	No workaround. Fix planned for future version
SNV VC	DRAGEN-25905	A single short target BED entry towards the end of a chromosome can cause a hang, for high depth samples.	Workaround is to either have more BED regions throughout the chromosome or increase bin memory
SNV VC	DRAGEN-26353	Contig names with disallowed colon character in VCFv4.2	Downstream tools such as htslib cannot process regions with contigs that contain colons. DRAGEN generates contig names as per input FASTA, or as per decoy list. This was in violation of VCF4.2 spec
System	DRAGEN-25358	dragen_hugepagctl conflicts with other programs that allocate hugepages	Customer attemts to run un-supported software on dragen server, causing a conflict in resources
UMI	DRAGEN-23614	Some UMI samples with ultra-high sequencing depths, can run into out-of-memory condition on on-site systems with 256GB RAM.	No workaround.
UMI	DRAGEN-25226	DRAGEN methylation app failures with UMI enabled	Certain Methylation UMI analyses fail on BSSH App

SW Installation Procedure

- Download the desired installer from the Illumina support website and unzip the package
- The archive integrity can be checked using: `./<DRAGEN 4.1.23 .run file> --check`
- Install the appropriate release based on your Linux OS with the command: `sudo sh <DRAGEN 4.1.23 .run file>`
- Please follow the installer instructions. Server power cycle may be required after installation, depending on the currently installed version. If an updated FPGA shell image needs to load from flash, this is only achieved with power cycle.

  - A power cycle is required when upgrading from v3.3.7 or older
  - A power cycle is required when downgrading to v3.3.7 or older
  - A power cycle is not required when upgrading from a release after v3.3.7
- Procedure to downgrade to v3.3.7 or older:
  - Requires the following three steps. The prior .mcs file needs to be flashed manually:
    - Install the prior release: `sudo sh <DRAGEN 3.3.7 .run file>`
    - `program_flash /opt/edico/bitstream/07*/*.mcs`
    - Power cycle

Release History

Revision	Release Reference	Originator	Description of Change
00	1092698	Cobus De Beer	Initial release